

Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment

Adrian Alan Pol^{1,5}, Virginia Azzolini^{1,4}, Gianluca Cerminara¹,
Federico De Guio^{1,6}, Giovanni Franzoni¹, Maurizio Pierini¹, Filip
Široký^{1,3}, Jean-Roch Vlimant² for the CMS Collaboration

¹ CERN, Meyrin, Switzerland

² California Institute of Technology, Pasadena, California, U.S.

³ Masaryk University, Brno, Czech Republic

⁴ Massachusetts Inst. of Technology, Cambridge, Massachusetts, U.S.

⁵ Université Paris-Saclay, Orsay, France

⁶ Texas Tech University, Lubbock, Texas, U.S.

E-mail: adrianalan.pol@cern.ch

Abstract. The certification of the CMS experiment data as usable for physics analysis is a crucial task to ensure the quality of all physics results published by the collaboration. Currently, the certification conducted by human experts is labor intensive and based on the scrutiny of distributions integrated on several hours of data taking. This contribution focuses on the design and prototype of an automated certification system assessing data quality on a per-luminosity section (i.e. 23 seconds of data taking) basis. Anomalies caused by detector malfunctioning or sub-optimal reconstruction are difficult to enumerate a priori and occur rarely, making it difficult to use classical supervised classification methods such as feedforward neural networks. We base our prototype on a semi-supervised approach which employs deep autoencoders. This approach has been qualified successfully on CMS data collected during the 2016 LHC run: we demonstrate its ability to detect anomalies with high accuracy and low false positive rate, when compared against the outcome of the manual certification by experts. A key advantage of this approach over other machine learning technologies is the great interpretability of the results, which can be further used to ascribe the origin of the problems in the data to a specific sub-detector or physics objects.

1. Introduction

Data certification (DC) process is the final step in the CMS Data Quality Monitoring (DQM) procedure. DC ensures the quality of data used for all physics results published by the CMS Collaboration. It is conducted by experts, who are trained to discover problems and pinpoint errors in the detector hardware based on assessment of dozens of histograms filled with certain critical quantities. The final certification flag is attributed comparing results to a predefined reference, representing the typical detector response during normal operation conditions. Using the histogram comparison, the knowledge of the LHC running conditions and the history of possible issues identified in the past, expert shifters evaluate problems. Further details on the infrastructure used for DQM are given in [1].

Current decisions by human experts are labor intensive and the histograms are integrated based on acquisition run basis. An acquisition run corresponds to a given setup both of the CMS detector and of LHC accelerator. Runs are denoted by integers, increasing with time. Their duration is varying from as little as few seconds to as much as several hours. One run could be a relatively long data taking interval. The work of pin-pointing the exact intervals of times affected by anomalous behavior can require further investigation and can require using non-event data as well. Hence, the certification flag can be inaccurate when transient problems throughout a run are overlooked or similarly useful data are thrown away from runs with malfunctions. A certification protocol based on shorter interval, using luminosity sections (LSs), is more desirable. Each run is divided into LSs, an interval corresponding to a fixed number of proton-beam orbits in the LHC and amounting to approximately 23 seconds, numbered progressively from 1 at the start of each run. Each LS can be identified uniquely by specifying the LS number and the run number. LS quality labels are already in place, obtained via application that monitors the powering and voltages delivered to the various subdetectors. Additional log messages (aggregated in [2]) also allow for fine grained analysis of the acquired monitoring data.

Machine learning (ML) methods open up the possibility to provide additional quality indicator in the current CMS DC procedure as the decision function can be learned directly from the copious archives of the past monitoring data and corresponding labels provided by experts. In the future, we hope to substantially filter the work required from human experts when the algorithm decision is certain and only invoke human judgment for questionable cases, as discussed in [3].

The remainder of the paper is organized as follows. Section 2 addresses challenges for applying ML in the context of DC. Section 3 discusses deep autoencoders in light of anomaly detection. Finally Section 4 presents the used dataset and the experimental setup and describes and discusses the results.

2. Challenges of CMS DC relevant to ML

2.1. Lumisection representation

Human detector experts make decisions regarding the data quality based on histograms. In case of an anomaly, histograms should show a considerable deviation from the nominal shape. To mimic the logic of current procedure we decided to represent each sample as a 2807 dimensional vector that is composed by five quantiles, mean and standard deviation of all distributions. We picked nearly all of the reconstructed particle collections (e.g. photons, muons, etc.) and represent all CMS subdetectors with a total of 401 physics observables (eg. energy, eta, phi etc.). All the distributions come from a dataset in so-called AOD [4] format. AOD format provides data for physics analysis in a convenient, compact format. It contains a copy of all the high-level physics objects, plus information sufficient to support typical analysis actions. We choose it as the best trade-off between the level of reconstruction (number of features needed to properly assess the quality of data for each LS) and the amount of information stored in those features.

2.2. Different event topologies

In the CMS experiment, the physics data is stored in different primary datasets (PDs). PDs are subsets of the event stream acquired by the CMS experiment grouped to satisfy constraints on the physics content, data processing and handling. Currently the DC process uses number of PDs tailored for the physics objective i.e. SingleMuon PD for *muons* or EGamma PD for *electrons*. For our primary study we have decided to use a dataset tailored for *jet* analysis which represents all CMS subdetectors since it contains every physics object and particle constituent needed for data quality certification. The proposed strategy has to be generic enough to be applicable for different PDs and in the future it is critical that the performance for all PDs is measured.

2.3. Class imbalance and sparsity of anomalies

Fortunately for the collaboration the data produced by the experiment is infrequently corrupted. Anomalies account for roughly 2% of the dataset. This makes classical ML supervised methods vulnerable to incomplete and inadequate representations of potential failures. Furthermore, due to changing root of the failure problems often tend to be novel. To account for this challenge we decided to explore a semi-supervised approach where a model learns the distribution of only one class (*good* in our case). In this manner we retain the potential to catch all the future unseen detector failures.

3. Anomaly detection using deep autoencoders

Autoencoder [5] (AE) is a parametric map from inputs to their representations, in the form of an artificial neural network. AEs are trained to perform an approximate identity mapping between their input and output layers. They consist of an encoder that takes an input and maps it to a usually lower-dimensional representation, and a decoder that tries to reconstruct the original input from the representation vector. The model should prioritize which aspects of the input should be distilled in order to learn useful properties of the data. Although it has been argued that, even for basic neural networks, most of the training is devoted to learning a compressed representation [6, 7], AEs are particularly suitable for anomaly detection. When trained on the good data, testing on unseen anomalous samples tend to yield sub-optimal representations and consequently decoder outputs. This indicates that a sample is likely generated by a different process, hence it should be flagged as problematic.

4. Experimental setup and results

In this experiment we use all 2807 features from all the LSs data recorded from June to October 2016 resulting in 163684 samples. During preprocessing we standardize the data by subtracting the mean and scaling the features to unit variance independently on each feature. Training (60%), validation (20%) and test (20%) rely on the quality labels provided by experts. We sort all samples chronologically and remove all the anomalies from training and validation sets.

We use architecture shown in Figure 1. The network has additional $L1$ kernel regularization on all the hidden nodes (10^{-5}). This sparsity constraint [8] penalizes the output of the hidden unit kernels and forces them to be close to zero. The hidden and output layers use parametric rectified linear units [9] as activations. We train the network with Keras [10] and TensorFlow [11] using the Adam optimizer [12] (with a learning rate of 0.0001) and early stopping mechanism monitoring validation dataset with patience set to 32 epochs. The network is instructed to minimize mean squared error between input X and the output \hat{X} vector: $\frac{1}{n} \sum_{i=0}^n (X_i - \hat{X}_i)^2$.

The final decision function is computed using mean squared error of the worst 100 reconstructed features (TOP100) to mirror human decision process. The difference between reference and recorded distributions is dominated by noise. Hence, experts pay attention only to significant deviations.

The final receiver operating characteristic (ROC) area under the curve (AUC) is 0.978, Figure 2 shows the error yield for samples in the held out test set. The visible waviness is associated with LHC instantaneous luminosity changes.

By examining the reconstruction error for each sample we can single out misbehaving features whose contribution to the overall error is high. Figure 3 shows a visualization for such investigation with grouped features (according to different physics objects).

5. Conclusions and Outlook

In this work we demonstrated that a semi-supervised anomaly detection using deep autoencoder based strategy can successfully produce certification flags with higher time granularity and with additional level of interpretability.

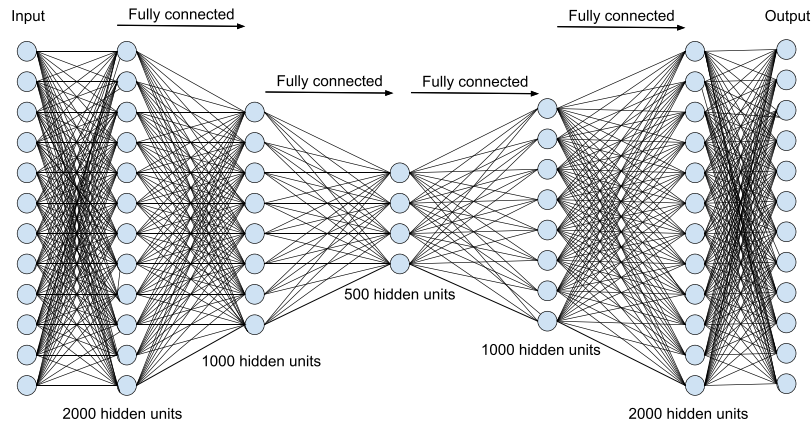


Figure 1. Proposed autoencoder architecture.

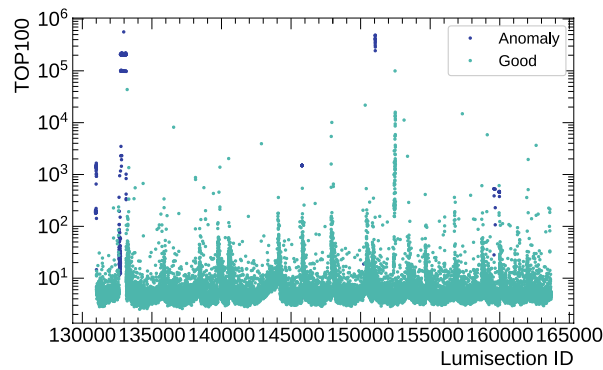


Figure 2. Mean of the worst 100 reconstructed features for each sample in the test set.

We will continue our research focusing on the hyperparams tuning, preprocessing and a strategy for automated model retraining (active learning). As part of the development we would like to address the instantaneous luminosity dependence of the reconstruction error. Finally, additional validation on different periods (i.e. 2017 data) and on different PDs would be necessary before integration in the production system.

Acknowledgments

We thank the CMS collaboration for providing the data set used in this study. We are thankful to the members of the CMS Physics Performance and Dataset project for useful discussions, suggestions, and support. We acknowledge the support of the CMS CERN group for providing the computing resources to train our models. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n° 772369).

References

- [1] M. Schneider, *The Data Quality Monitoring Software for the CMS experiment at the LHC: past, present and future*, in *this conference* (2018)
- [2] V. Rapsevicius et al., *CMS Run Registry: Data certification bookkeeping and publication system*, in *IOP Conf. Ser J Phys Confer Ser* (2011), **331**, p. 042038

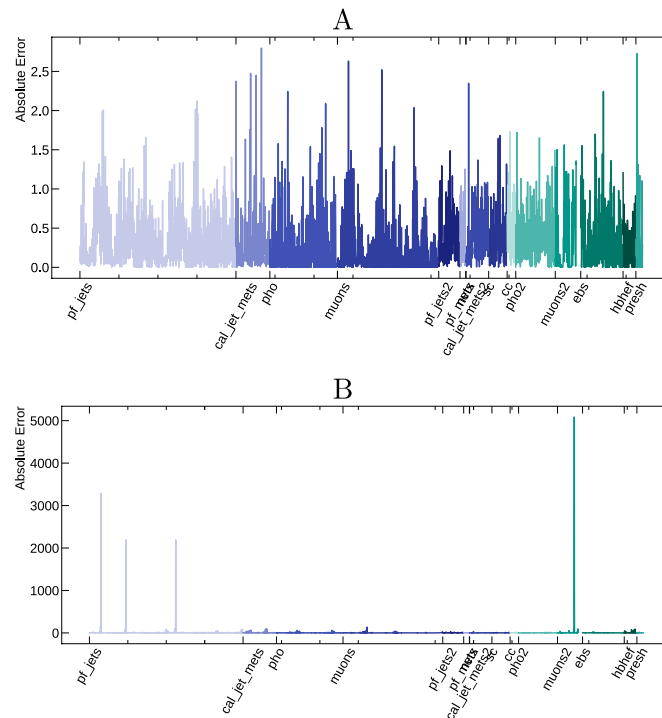


Figure 3. Reconstruction error for two samples for each feature. Different colors represent features linked to different physics objects. For a good sample (A) we can expect similar amplitude across all objects with small absolute scale. Anomalous samples (B) have clearly visible peaks for problematic features (muons - *muons2* and jets - *pf jets*).

- [3] M. Borisyak, F. Ratnikov, D. Derkach, A. Ustyuzhanin, *Towards automation of data quality system for CERN CMS experiment*, in *IOP Conf. Ser J Phys Confer Ser* (2017, doi: 10.1088/1742-6596/898/9/092041), **898**, p. 092041
- [4] M. Della Negra, L. Foà, A. Hervé, A. Petrilli, Tech. Rep. CERN/LHCC-2005-023, CMS computing (2005)
- [5] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning (pages 499-523)* (MIT Press, 2016)
- [6] N. Tishby, N. Zaslavsky, *Deep learning and the information bottleneck principle* (2015). arXiv:1503.02406
- [7] R. Shwartz-Ziv, N. Tishby, *Opening the black box of deep neural networks via information* (2017). arXiv:1703.00810
- [8] M. Ranzato, C. Poultney, S. Chopra, Y. LeCun, *Efficient Learning of Sparse Representations with an Energy-based Model*, in *Proceedings of NIPS* (2006), pp. 1137–1144
- [9] K. He, X. Zhang, S. Ren, J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, in *Proceedings to ICCV* (2015), pp. 1026–1034
- [10] F. Chollet et al., *Keras*, <https://keras.io> (2015)
- [11] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., *Tensorflow: a system for large-scale machine learning.*, in *OSDI* (2016), **16**, pp. 265–283
- [12] D. Kingma, J. Ba, *Adam: A method for stochastic optimization* (2014). arXiv:1412.6980